Clustering Clauses for High-Level Relation Detection: An Information-theoretic Approach

Samuel Brody

School of Informatics
University of Edinburgh
s.brody@sms.ed.ac.uk

ACL 2007, Prague
Introduction

Approach
- Structural Representation
- Generalization
- Information Bottleneck Clustering

Evaluation
- via WordNet
- Relation Detection Task

Related Work

Conclusions
Outline

1. Introduction

2. Approach
   - Structural Representation
   - Generalization
   - Information Bottleneck Clustering

3. Evaluation
   - via WordNet
   - Relation Detection Task

4. Related Work

5. Conclusions
Motivation

High-Level Relations
Relations involving whole sentences

Useful for: Question Answering, Discourse Analysis, Summarization, Paraphrasing, ...

Challenges
- sparseness of data
- data representation
- how to generalize correctly
- adaptable to domain and task
Goal

Training Data

The king condemned the cruel foe.
The newly-elected president denounced his opponent.
The commanders severely criticized their rival.

Generalized Representation

Test Data

The resigning mayor blamed his competitors for his failure.
The military commander has publicly denounced his foes.
1 Introduction

2 Approach
   • Structural Representation
   • Generalization
   • Information Bottleneck Clustering

3 Evaluation
   • via WordNet
   • Relation Detection Task

4 Related Work

5 Conclusions
subject, verb and object usually convey the essence of the sentence.

common to most sentences.
Generalization

Clustering - Advantages

- represents a general concept via a single unit
- unsupervised
- domain and corpus specific
- can vary the granularity
Training Data

The king condemned the cruel foe.
The newly-elected president denounced his opponent.
The commanders severely criticized their rival.

Generalized Representation

Test Data

The resigning mayor blamed his competitors for his failure.
The military commander has publicly denounced his foes.
Training Data

The king condemned the cruel foe.
The newly-elected president denounced his opponent.
The commanders severely criticized their rival.

Clustered Clause

(subject cluster) 
- king
- president
- mayor
- chief
- commander

(verb cluster) 
- denounce
- criticize
- blame
- condemn

(object cluster) 
- rival
- competitor
- foe
- opponent

Test Data

The resigning mayor blamed his competitors for his failure.
The military commander has publicly denounced his foes.
Information Bottleneck Principle [Tishby et al. 1999]

Seeks to compress the values of one variable with regard to information about another (target) variable.

<table>
<thead>
<tr>
<th>S \ V</th>
<th>tell</th>
<th>scratch</th>
<th>drink</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>John</td>
<td>4</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>cat</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>man</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

This clustering preserves the highest Mutual Information with regard to the associated verbs.

<table>
<thead>
<tr>
<th>Clustered S \ V</th>
<th>tell</th>
<th>scratch</th>
<th>drink</th>
</tr>
</thead>
<tbody>
<tr>
<td>{dog, cat}</td>
<td>0</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>{John, man}</td>
<td>10</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>
Clustering Setup

Two phase procedure:
- cluster subjects and objects according to verbs
- cluster verbs according to subject and object clusters

![Diagram of clustering setup with phases 1 and 2]
Experimental Setup

- Reuters Corpus (English, release 2000)
- parsed using Minipar [Lin (1994)]
- 3,153 subjects, 3,312 objects and 1,716 verbs
- 2.8 million clauses
- 200 subject & object clusters, 100 verb clusters
Cluster Examples

Subject Cluster:
troop, marine, military, bodyguard, peacekeeper, navy, slorc, guard, fighter, police, soldier, rebel, squad, taleban, patrol, force, militia, army, tank, policemen

Verb Cluster:
mull, study, envision, reconsider, analyse, assess, inherit, pursue, disclose, clarify, contemplate, analyze, discuss, examine, deem, recommend, decide, manage, specify, prefer, review, weigh, monitor, determine, evaluate, define, intend, explore, consider, assume

Object Cluster:
burn, wound, harm, embarrassment, defeat, setback, injury, damage, casualty, stroke, blow
Outline

1. Introduction

2. Approach
   - Structural Representation
   - Generalization
   - Information Bottleneck Clustering

3. Evaluation
   - via WordNet
   - Relation Detection Task

4. Related Work

5. Conclusions
compare average pairwise similarity in our clusters (blue) to random clustering of the same words (red)

similarity measured using three independent methods from the WordNet::Similarity package [Pedersen et al. (2004)]
Why a task?
- provides pragmatic evaluation for clustering
- demonstrates application of clustered model

Why relation detection?
- natural environment involving whole sentences
- general and important for many applications
Recognizing Relations

- look for pairs of sentences, not too far apart, containing a common noun (anchor).
- group together similar pairs: containing words from the same cluster in each position.

**Example: Subject-Subject Anchor**

- “Peter twisted his ankle” ... “Peter let out a shout”
- “Ted broke his arm” ... “Ted emitted a yell”
- “Jane cut her finger” ... “Jane let out a shriek”

- do these groups represent a common relation?
detected groups are ranked by statistical significance
those with $p$-value $< 0.05$ are retained

<table>
<thead>
<tr>
<th>Anchoring System</th>
<th>Number of Relations Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>697</td>
</tr>
<tr>
<td>Subject-Subject</td>
<td>428</td>
</tr>
<tr>
<td>Object-Object</td>
<td>291</td>
</tr>
<tr>
<td>Subject-Object</td>
<td>180</td>
</tr>
<tr>
<td>Object-Subject</td>
<td>178</td>
</tr>
</tbody>
</table>
Relation Examples

- storm, hit, coast
- quake, hit, northwest
- earthquake, hit, city
- cyclone, near, coast

- storm, cause, damage
- quake, cause, casualty
- earthquake, cause, damage
- cyclone, cause, damage

- sun, report, earnings
- xerox, report, earnings
- microsoft, release, result

- earnings, beat, expectation
- earnings, beat, forecast
- result, beat, forecast
Human Evaluation

For the human judges, clause stubs were transformed to sentences: e.g. “cyclone, near, coast” → “The cyclone neared the coast.”

<table>
<thead>
<tr>
<th>Anchoring</th>
<th>Instances</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3.5</td>
<td>2.6</td>
</tr>
<tr>
<td>O-S</td>
<td>3.9</td>
<td>2.9</td>
</tr>
<tr>
<td>O-O</td>
<td>3.5</td>
<td>2.7</td>
</tr>
<tr>
<td>S-O</td>
<td>3.5</td>
<td>2.6</td>
</tr>
<tr>
<td>S-S</td>
<td>3.2</td>
<td>2.4</td>
</tr>
</tbody>
</table>

- judges asked to rate groups of sentence pairs
- rating from 1-5 signifying common relationship
- baseline: groups of unconnected sentence pairs

System output is significantly better than baseline
Outline

1. Introduction

2. Approach
   - Structural Representation
   - Generalization
   - Information Bottleneck Clustering

3. Evaluation
   - via WordNet
   - Relation Detection Task

4. Related Work

5. Conclusions
Related Work

- clustering individual words
  
  [Hindle (1990); Pereira et al. (1993); Lin (1998)]

- sentence similarity
  
  [Lin and Pantel (2001)]

- high-level relations without generalization
  
  [Chklovski and Pantel (2004); Torisawa (2006)]
Outline

1. Introduction

2. Approach
   - Structural Representation
   - Generalization
   - Information Bottleneck Clustering

3. Evaluation
   - via WordNet
   - Relation Detection Task

4. Related Work

5. Conclusions
Our Method

Contributions
- generalized representation through clusters
- unsupervised $\rightarrow$ tailorable to a domain
- helps overcome sparseness

Global approach
- general approach for detecting relations in complex sequences (text, biological data, DB access ... )
  1. use local structure and information for representation & clustering
  2. detect high-level relations using clustered representation
Future Work

- expansion of the clause model
- hierarchical clustering for higher precision
- exploring different domains
- other relation patterns
Thank You!
Baseline Examples

Greenspan saw the pressure. – Greenspan rocked the market.
The union approved the strategy. – The union mortgaged the property.
The plant made helicopters. – The plant resumed production.

The bill included a provision. ... The bill included a proposal.
Opel sold the trucks. ... Opel hired engineers.
The company recorded growth. ... The company announced a plan.
## Human Evaluation

<table>
<thead>
<tr>
<th>Anchoring System</th>
<th>Instances Score</th>
<th>Baseline Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3.5</td>
<td>2.6</td>
</tr>
<tr>
<td>O-S</td>
<td>3.9</td>
<td>2.9</td>
</tr>
<tr>
<td>O-O</td>
<td>3.5</td>
<td>2.7</td>
</tr>
<tr>
<td>S-O</td>
<td>3.5</td>
<td>2.6</td>
</tr>
<tr>
<td>S-S</td>
<td>3.2</td>
<td>2.4</td>
</tr>
</tbody>
</table>

- human tendency for false positive
- tendency varies with type of pattern
## Human Evaluation

<table>
<thead>
<tr>
<th>Anchoring System</th>
<th>Instances Score</th>
<th>Baseline Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3.5</td>
<td>2.6</td>
</tr>
<tr>
<td>O-S</td>
<td>3.9</td>
<td>2.9</td>
</tr>
<tr>
<td>O-O</td>
<td>3.5</td>
<td>2.7</td>
</tr>
<tr>
<td>S-O</td>
<td>3.5</td>
<td>2.6</td>
</tr>
<tr>
<td>S-S</td>
<td>3.2</td>
<td>2.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anchoring System</th>
<th>Number of Patterns Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>697</td>
</tr>
<tr>
<td>Subject-Subject</td>
<td>428</td>
</tr>
<tr>
<td>Object-Object</td>
<td>291</td>
</tr>
<tr>
<td>Subject-Object</td>
<td>180</td>
</tr>
<tr>
<td>Object-Subject</td>
<td>178</td>
</tr>
</tbody>
</table>

- human tendency for false positive
- tendency varies with type of pattern
- variation may be due to expected pattern frequency


